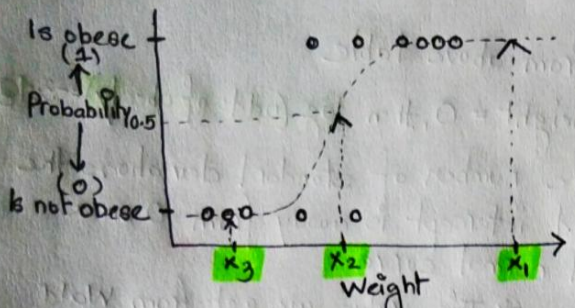
LOGISTIC REGRESSION

→ Logistic regression predict the categories (categorical values) instead of continuous values. Logistic regression is a specific type of Generalized Linear Model (GLM).

→ Example - If the mice are obese or not.



→ Logistic regression fits "S" shaped logistic function.

→ The curve goes from 0 to 1 that means the curve tells us what the probability that a mouse is obese based on its weight.

→ If we weight a heavy mouse (x_1), there is high probability new mouse is obese.

→ If we weight an intermediate mouse (x_2), then there is only a 50% chance that the mouse is obese.

→ If we weight a small mouse (x_3), then there is only small probability that a light mouse is obese.

→ Although logistic regression tells the probability that a mouse is obese or not, it is used for classification.

→ For example, if the probability a mouse is obese is > 50%, then we will classify it as obese, otherwise we will classify it as "not obese".

→ Logistic regression can work with continuous data (like weight and age) and discrete data (like genetic type and astrological sign) both.

Coefficients →

→ The y axis in logistic regression is transformed from the "probability of obesity" to the "log(odds of obesity)" so just like y-axis in linear regression, it can go from -infinity to +infinity.

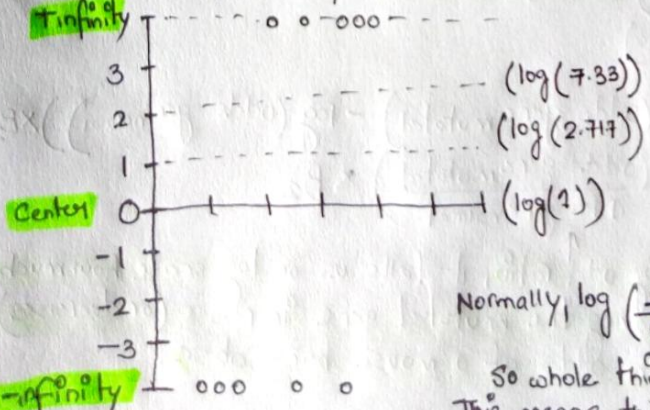
$$\log(\text{odds of obesity}) = \log\left(\frac{p}{1-p}\right)$$

p in this case is the probability of a mouse being obese and corresponds a value on the old y axis between 0 and 1.

→ Suppose $p = 0.5$, $\log\left(\frac{0.5}{1-0.5}\right) = \log\left(\frac{0.5}{0.5}\right) = \log(1)$

$\log(1) = 0$.

The center of new y-axis.



Suppose $p = 0.731$, $\log\left(\frac{0.731}{0.269}\right) = \log(2.717) = 1$.

$p = 0.88$, $\log\left(\frac{0.88}{0.12}\right) = \log(7.33) = 2$.

$p = 0.95$, $\log\left(\frac{0.95}{0.05}\right) = \log(19) = 3$.

All the points are at $p = 1$,

$\log\left(\frac{1}{1-1}\right) = \log\left(\frac{1}{0}\right) = \infty$

Normally, $\log\left(\frac{1}{0}\right) = \log(1) - \log(0)$, $\log(0)$ is define as negative infinity

$= \text{something} - \text{negative infinity} = \text{positive infinity}$

So whole thing is equal to positive infinity

This means the original sample of obese are at positive infinity on y-axis

As a result, probability of 0.5 to 1 converted to 0 to +infinity. Similarly do for negative side.

Logistic Output → NUMERICAL VARIABLE →

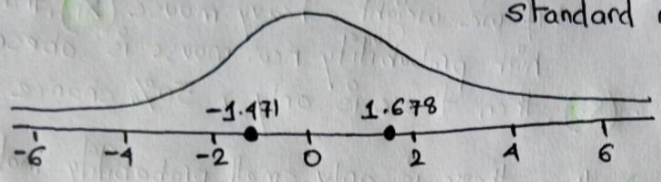
Coefficients :

	Estimated	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.476	2.364	-1.471	0.1414
weight	1.825	1.088	1.678	0.934

$\boxed{Y = -3.48 + 1.83 \times \text{weight}}$ → Derived from above table

-3.476 → ~ -3.48 → Y intercept when weight = 0, then log (odds of obesity) = -3.48

Zvalue = $\frac{\text{Estimate}}{\text{Std. Error (Standard error)}}$ = It is the number of standard deviations the estimated intercept is away from 0 on a standard normal curve.



→ This value we get from Wald test.

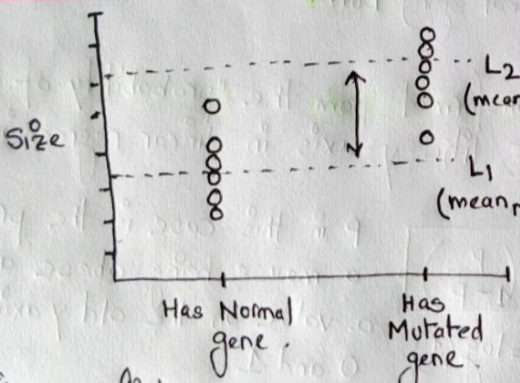
→ Since the estimate is less than 2 standard deviations away from 0, we know its not statiscally significant. And this confirm by large p value.

→ 1.825 → It means for every one unit of weight gained, the log (odds of obesity) increased by 1.825

→ Now we understood coefficient of continuous variable, now time for categorical variable.

Categorical Variable → We perform t test for categorical value.

This is how t-test works →



Step 1 → Plot the points.

Step 2 → Fit two lines to the data. Line represent the mean size of normal gene (L_1) & mutated gene (L_2)

Step 3 → Equation formula,

$\text{size} = \text{mean}_{\text{normal}} \times B_1 + (\text{mean}_{\text{mutant}} - \text{mean}_{\text{normal}}) \times B_2$

When we do a t-test this way, we basically testing to see if coefficient $(\text{mean}_{\text{mutant}} - \text{mean}_{\text{normal}})$ is equal to zero.

So in logistic we find,

$\log(\text{odds gene}_{\text{normal}}) = -1.55$ (suppose)

$\log(\text{odds gene}_{\text{mutated}}) = 0.85$ (suppose)

$\text{size} = \log(\text{odds gene}_{\text{normal}}) \times B_1 + (\log(\text{odds gene}_{\text{mutated}}) - \log(\text{odds gene}_{\text{normal}})) \times B_2$

$= \log(\text{odds gene}_{\text{normal}}) \times B_1 + \log\left(\frac{\text{odds gene}_{\text{mutated}}}{\text{odds gene}_{\text{normal}}}\right) \times B_2$

$= \log\left(\frac{2}{2}\right) \times B_1 + \log\left(\frac{7/3}{2/3}\right) B_2$

$= -1.5\, B_1 + 2.35\, B_2$.

log odd ratio, it tells us on log scale how much having the mutated gene increases (or decreases) the odds of a mouse being obese.

In output we see,

Coeff	Estimate	Std Error	zvalue	pvalue
Intercept	-1.54	0.7817	-1.929	0.559
geneMutant	2.35	1.048	1.0429	2.225

Suppose the output is as follows →

glm(formula = hd ~ a+b+c, family = "binomial", data = data)

Deviance Residual :

Min	1Q	Median	3Q	Max
-1.2	-1.27	-0.776	1.08	1.61

Residual deviance looks good they are close to being centered on 0 and are roughly symmetrical.

Coefficients :

	Estimate	Std Error	zvalue	Pvalue.
(Intercept)	1.74	0.99	1.75	0.79
A	-0.39	0.32	-1.23	0.21
B	-0.12	0.006	-2.057	0.03
C	0.18	0.008	-3.08	0.04

Null deviance : 234.67 on 188 degree of freedom

Residual deviance : 227.38 on 186 degree of freedom.

AIC : 213.12.

Number of Fisher Scoring iterations : 4.

→ If the proposed model has a good fit, the deviance will be small.

→ If the model has bad fit, deviance will be high.

→ RESIDUAL DEVIANCE = 227.38 on 186 DF

when the model has included A, B and C variable, then the deviance is residual deviance which is lower (227.38) than null deviance (234.67). Lower value of residual deviance points out that the model has become better when it has included 3 variables (A, B, C).

→ Null deviance = 234.67 on 188 DF

when the model includes only one intercept, then the performance of the model is governed by NULL deviance.

→ Fisher Scoring Iteration : 4

- In short it says model needed 4 iterations to perform the fit.
- The algorithm looks around to see if the fit would be improved by using different estimates. If it improves then it moves in that direction and then fits the model again. The algorithm stops when no significant additional improvement can be done.
- In this case, best model acheived in 4 iterations.

→ AIC = 213.12

- AIC is Alkaline Information Criterion.
- This is useful when we have more than one model to compare goodness of fit.
- It is maximum likelihood which penalize overfitting.
- Lower AIC of model is better than model having higher AIC.

→ P value → If p-value is less than 0.05 then the variable are statistically significant.

So variable A is not significant (0.21) and variable B (0.03), variable C (0.04) both are statistically significant.

→ We can also calculate MacFadden's Pseudo R^2

Suppose Pseudo $R^2 = 0.55$. This can be interpreted as overall effect size.

→ And we can calculate a p-value for pseudo R^2 using chi-square distribution.

Suppose p-value = 0. So pvalue < 0.05 so R^2 is now with damn luck.